Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 177 (2018) 405–425

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Stable matching mechanisms are not obviously strategy-proof ☆

Itai Ashlagi [a,*], Yannai A. Gonczarowski [b,c,*]

[a] *Management Science & Engineering, Stanford University, United States of America*
[b] *Einstein Institute of Mathematics, Rachel & Selim Benin School of Computer Science & Engineering, and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel*
[c] *Microsoft Research, Israel*

Received 30 July 2017; final version received 14 March 2018; accepted 9 July 2018

Abstract

Many two-sided matching markets, from labor markets to school choice programs, use a clearinghouse based on the applicant-proposing deferred acceptance algorithm, which is well known to be strategy-proof for the applicants. Nonetheless, a growing amount of empirical evidence reveals that applicants misrepresent their preferences when this mechanism is used. This paper shows that no mechanism that implements a stable matching is *obviously strategy-proof* for any side of the market, a stronger incentive property than strategy-proofness that was introduced by Li (2017). A stable mechanism that is obviously strategy-proof for applicants is introduced for the case in which agents on the other side have acyclical preferences.

JEL classification: D4; D47

Keywords: Stable matching; Obviously strategy-proof; Obvious strategy-proofness; Matching; Mechanism

☆ This paper greatly benefited from discussions with Sophie Bade, Elchanan Ben-Porath, Shengwu Li, Jordi Massó, Muriel Niederle, Noam Nisan, Assaf Romm, and Peter Troyan. Itai Ashlagi is supported by NSF grant SES-1254768. Yannai Gonczarowski is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities; his work is supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 249159, by ISF grant 1435/14 administered by the Israeli Academy of Sciences, and by Israel-USA Bi-national Science Foundation (BSF) grant number 2014389.
* Corresponding authors.
E-mail addresses: iashlagi@stanford.edu (I. Ashlagi), yannai@gonch.name (Y.A. Gonczarowski).

https://doi.org/10.1016/j.jet.2018.07.001

1. Introduction

A number of labor markets and school admission programs that can be viewed as two-sided matching markets use centralized mechanisms to match agents on both sides of the market (or agents on one side of the market and objects on the other side of the market). One important criterion in the design of such mechanisms is stability (Roth, 2002), requiring that no two agents, one from each side of the market, prefer each other over the partners with whom they are matched. Another highly desired property is strategy-proofness, which alleviates agents' incentives to behave strategically.[1]

Indeed, many clearinghouses have adopted in recent years the remarkable deferred acceptance (DA) mechanism (Gale and Shapley, 1962),[2] which finds a stable matching and is strategy-proof for one side of the market, namely the proposing side in the DA algorithm (Dubins and Freedman, 1981).[3,4] Interestingly, although participants are advised that it is in their best interest to state their true preferences, empirical evidence suggests that a significant fraction nonetheless attempt to strategically misreport their true preferences (Hassidim et al., 2017); this was observed in experiments (Chen and Sönmez, 2006), in surveys (Rees-Jones, 2016), and in the field (Hassidim et al., 2016; Shorrer and Sóvágó, 2017). This paper asks whether one can implement the deferred acceptance outcome via a mechanism whose description makes its strategy-proofness more apparent. Toward this goal, we adopt the notion of *obvious strategy-proofness*, an incentive property introduced by Li (2017) that is stronger than strategy-proofness.

Li (2017) formulated the idea that it is "easier to be convinced" of the strategy-proofness of some mechanisms over others. He introduces, and characterizes, the class of *obviously strategy-proof* mechanisms. He shows that, roughly speaking, obviously strategy-proof mechanisms are those whose strategy-proofness can be proved even under a cognitively limited proof model that does not allow for contingent reasoning.[5] In his paper, Li studies whether various well-known auction and assignment mechanisms with attractive revenue or welfare properties for one side of the market can be implemented in an obviously strategy-proof manner. Whether one may implement stable matchings in an obviously strategy-proof manner remained an open problem.

For the purpose of this paper, we adopt the Gale and Shapley (1962) one-to-one matching market with men and women to represent two-sided matching markets; our main results naturally extend to many-to-one markets such as labor markets and school choice programs. When women's preferences over men are perfectly aligned, the unique stable matching may be recovered via serial dictatorship, where men, in their ranked order, choose their partners. In this case, a sequential implementation of such serial dictatorship is obviously strategy-proof. (This follows from Li, 2017, who shows that in a two-sided assignment market with agents and objects, serial

[1] See also Pathak and Sönmez (2008), which finds that non-strategy-proof mechanisms favor sophisticated players over more naïve players.

[2] Examples include the National Resident Matching Program (Roth, 1984), as well as school choice programs in Boston (Abdulkadiroğlu et al., 2005) and New York (Abdulkadiroğlu et al., 2009) (see also Abdulkadiroğlu and Sönmez, 2003).

[3] This mechanism is also approximately strategy-proof for all participants in the market (Immorlica and Mahdian, 2005; Kojima and Pathak, 2009; Ashlagi et al., 2017).

[4] Indeed, removing the incentives to "game the system" was a key factor in the city of Boston's decision to replace its school assignment mechanism in 2005 (Abdulkadiroğlu et al., 2006).

[5] For instance, this notion separates sealed-bid second-price auctions from ascending auctions (where bidders only need to decide at any given moment whether to quit or not) and provides a possible explanation as to why more subjects have been reported to behave insincerely in the former than in the latter (Kagel et al., 1987).